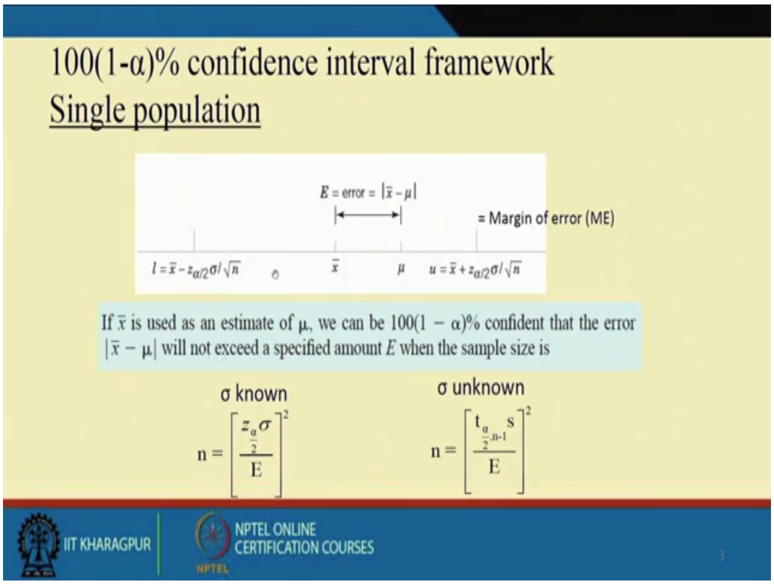
Design and Analysis of Experiments
Prof. Jhareswar Maiti
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture - 12
Determination of Sample Size

Welcome. In this lecture we will discuss; how to determine the sample size for an experimentation. It is a very important concept, because sample size plays in significant role in accomplish of the measures. We will cover this in 2 different lectures. In first lecture now, we will discuss with reference to single population mean test and difference between two population mean test. So, contains of presentation today. We will see that how the sample size can be calculated using confidence intervals framework and hypothesis testing framework.

(Refer Slide Time: 01:16)

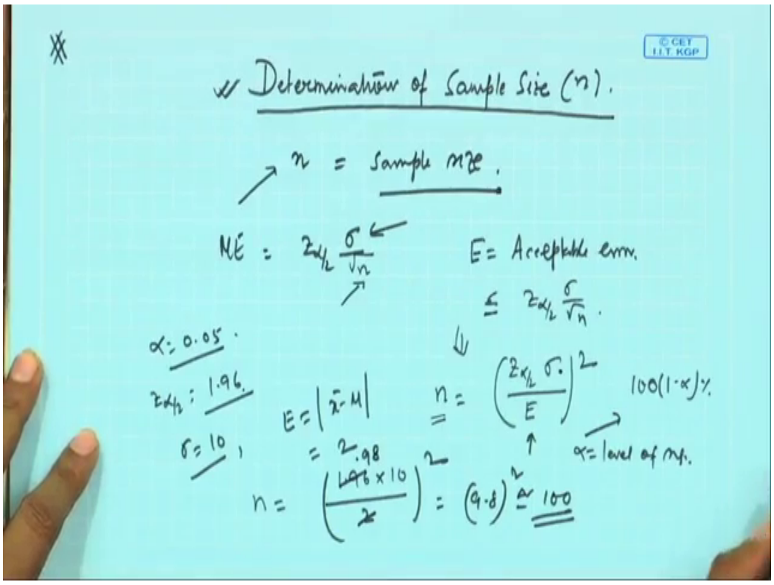


You see this figure. Here you just consider in normal population single population with min mu and standard deviations sigma.

Suppose you have conducted one sam you have collected one sample of size n. Now you are interested to know the confidence interval of mu from these sample calculations. So, x bar is the estimate of mu, and x bar is kept here. And then you have found out 100 into 1 minus alpha percent confidence which will be from between l to u. And we have seen earlier l equal to x bar minus z alpha by 2 sigma by root n and u is x bar plus z alpha by 2 sigma by root n. Now this plus minus sigma or sigma f z alpha by 2 sigma by root n means from x bar to u as well as from x bar to l this is known as margin of error.

Now, what we assume that the mu lies in between this interval with a confidence of 100 into 1 minus alpha percent. We do not know that what is the distance between x bar and mu, mu is somewhere. So, let us the permissible error is E which is x bar minus mu absolute value. Now we should calculate n in such a manner that, that these error lie within the margin of error.

(Refer Slide Time: 03:39)



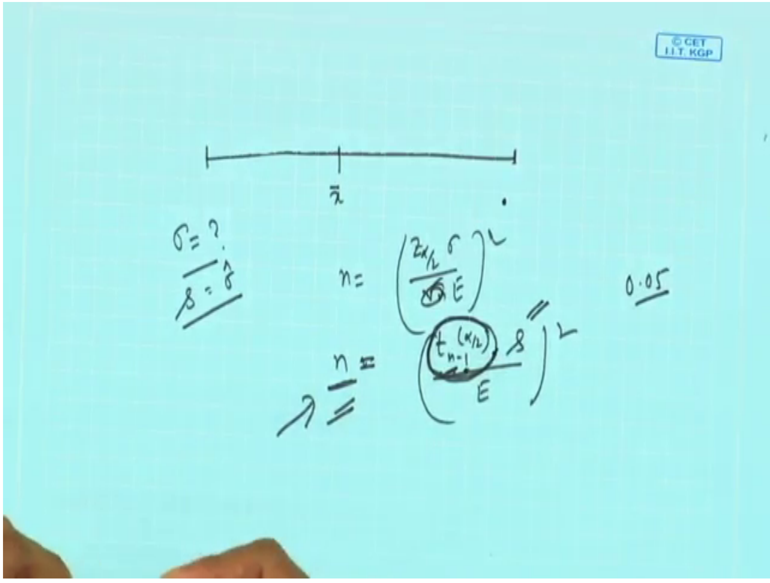
That means, the maximum error quantity will be ME. So, that is the concept. So, if I say then what is ME in this case, z alpha by 2 sigma by root n. You see sigma is the population standard deviation; it is fixed value you do not have n control on it. Now where you have control you have control on n. So, if we say that E which is the acceptable error equal to it will be maximum it will be less than equal to z alpha by 2 sigma by root n. And using this you can find out n equal to z alpha by 2 sigma by E whole square.

So, if you know sigma, if you then for a particular confidence interval for a par particular confidence interval means 100 into 1 minus alpha percent, or for a particular level of significance alpha, you will be avail to compute n; provided you know the E. Suppose let alpha equal to 0.05, then z alpha by 2 will be 1.96. And let us sigma equal to 10, and suppose E the accept equal to that x bar minus mu, this will acceptable one is this is 2.

So, then what will be n? N will be 1.96 into 10 divided by 2 square. That mean this is 9.98, but and this into 10. So, 9.8 square; so this will be if I say that this is around 10, and it is almost 100. So, this is the way you calculate sample size.

Now, sometimes what happened in this confidence interval approach, you will you do not know what is the sigma value.

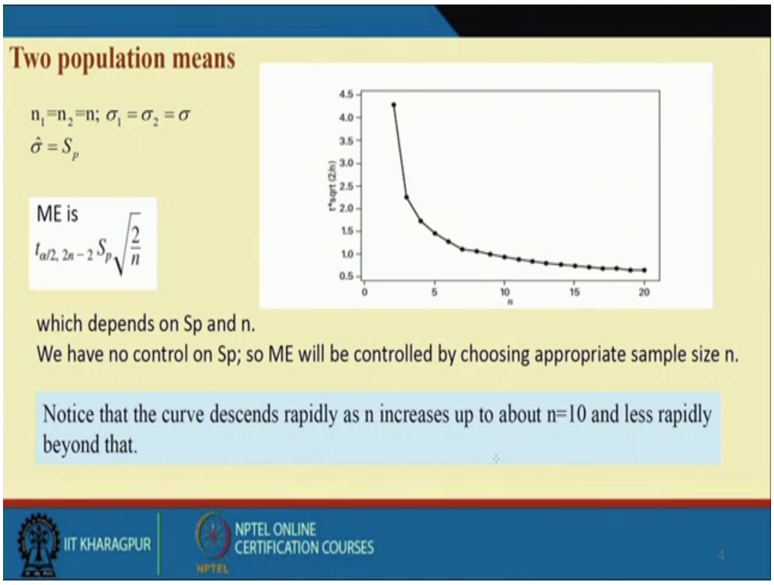
(Refer Slide Time: 06:12)



In that case you use s as the estimate of sigma, and then what you require you cannot use (Refer Time: 06:28) this z distribution, you will be using t distribution. And then n instead of z alpha by 2 sigma by root n sigma by E square it will be n will be t n minus 1 alpha by 2 into s by E whole square, where s is the sample standard deviation. So, essentially if even if you say alpha equal to 0.05, then depending on this n suppose you choose a particular n t s degrees of freedom will be difference. Second n like this you will get (Refer Time: 07:16) will be difference. So, you have to find out that n which ultimately satisfy the satisfy the condition, then it will be

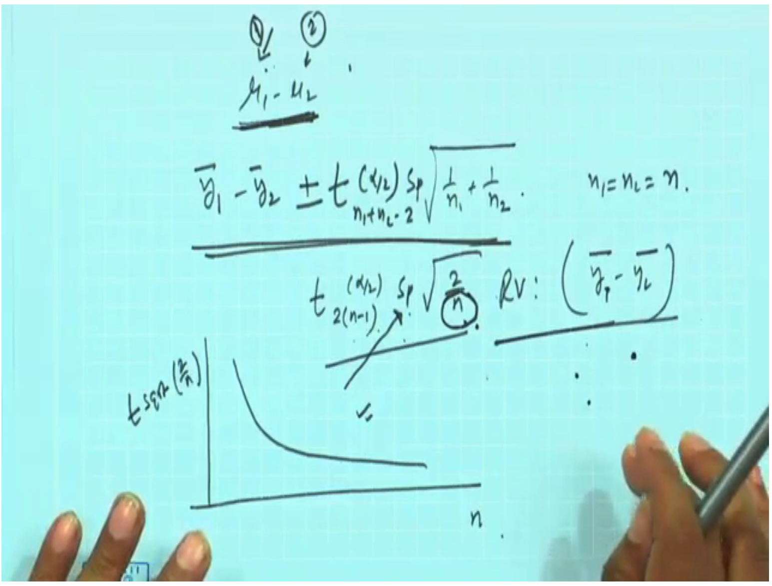
So, as t is t n minus 1, this quantity again dependent on n. So, the formula what is written as n equal to t alpha by 2 n minus 1 s square by E; this will only give you for a particular n what is this? You choose n then t alpha by 2 n s by E, but you have to go for trial and error method.

(Refer Slide Time: 07:55)



This is what we have shown here for t distribution. Here instead of single population, we are use two populations. And our test is or we are interested in the confidence interval for the difference between two population means like mu 1 minus mu 2.

(Refer Slide Time: 08:19)

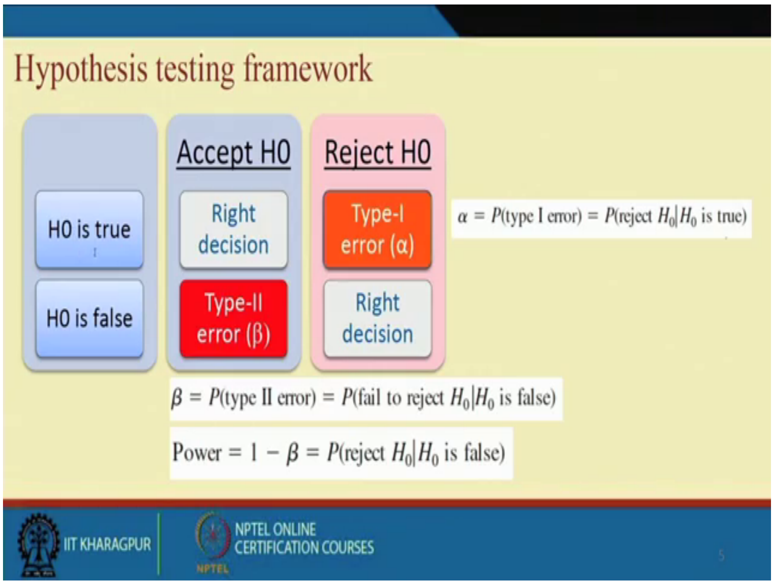


So, mu 1 is the population for the population mean for the characteristics y. And this is for first population this is for second population. And this will be the difference. So, we want to know the difference between two population means, whether they are different or not. So, we suppose you want the confidence interval for this. Then you all know you have seen earlier this will be nothing but y 1 bar minus y 2 bar, minus we can write t n 1 plus n 2 minus 2 alpha by 2, and then root over 1 by n 1 plus 1 by n 2 and s p will be there. So, this is plus minus this is the interval.

Now, if n 1 equal to n 2, equal to n, then this quantity become t 2 into n minus 1 alpha by 2 s p root over 2 by n t alpha by 2 to n minus 1 s p root over 2 by n. So, here what happened? You see that what is the countable? S p is uncountable for you because this is the sample pulled variance. So, we have n which can be controlled. So, this n, now if we plot; that is, t square root of 2 by n and this side n will get a curve. The curve will be something like this for this shown here. You see we have we have plotted this versus this. And ultimately you see that with n increases, this quantity decreases, and at around 10 or 11 after that there is not much decrease. So, you can assume that the n E 10 is a good sample size.

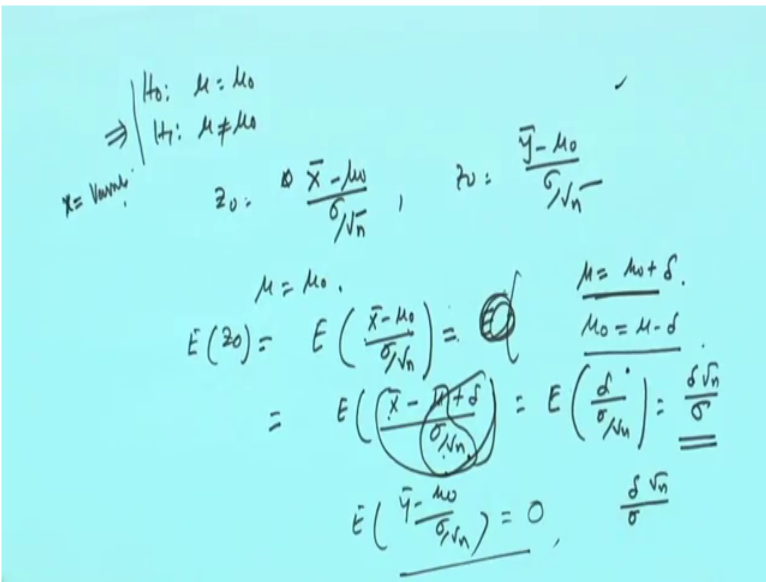
So, notice that the curve descends rapidly as n increase up to about 10 and less rapidly beyond that. So, you can say n equal to 10 is a reasonable good sample size; so this is what is our confidence interval approach. So, first I have giving you for single population and for two population main differences. If it is if the random variable y 1 bar minus y 2 bar this called this giving situation, it follows normal distribution. It is easier, or if it follows t distribution, then you have to choose n. Then the degrees are using that corresponding degrees of freedom find out the t values. So, you require iterative method whether that is single population or two population case. And then you find out that n you consider that n which gives you the less t square root of 2 by n.

(Refer Slide Time: 12:05)



Now, we will discuss another approach called hypothesis testing approach. What you do in hypothesis testing? You assume some null hypothesis and alternate hypothesis, and under null hypothesis you test using appropriate statistic and statistical and sampling distribution. And then you reject the null hypothesis depending on the test result or you failed to reject the null hypothesis. As a result, what happens? There will be reverse 2 are favourable and 2 are unfavourable by unfavourable. What I mean to say that they we commits certain errors and that is what is the decision framework for hypothesis testing.

(Refer Slide Time: 12:57)

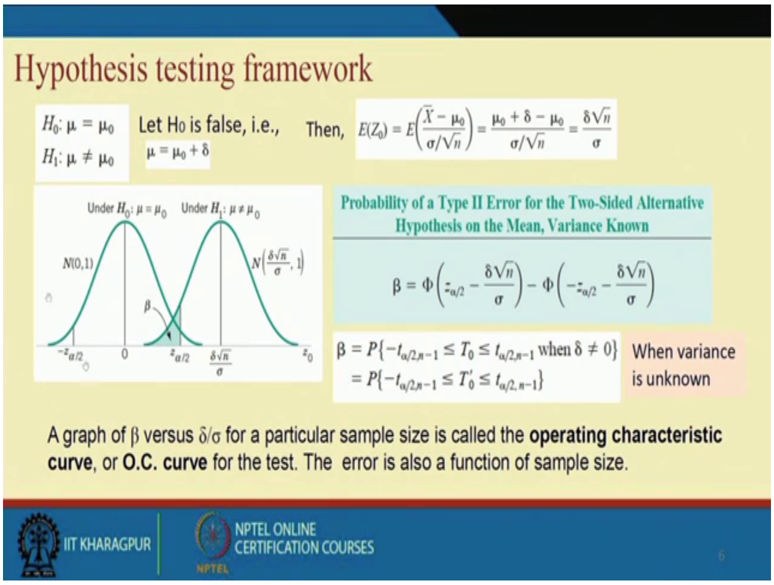


Now, you see the suppose H 0 for let it be a single population. H 0 we are saying mu equal to m 0 and H 1 we are saying mu not equal to mu 0. So, this is what you have already you seen earlier. Now the situation may be H 0 is actually true and H 0 actually false. But you would test may accept a false H 0 or may reject a correct H 0. So, that is the situation here. If H 0 is true and your test accepted that it right decision, if H 0 is false test accepted it is again right decision. But if H 0 is true and you are rejecting null H 0, then it is an error for type one error which probability is alpha. So, alpha is probability of committing type one error that mean probability reject H 0, H 0 is true.

Similarly, another one H 0 is false, but you are accepting it this is known as B type 2 error and beta is the probability of type 2 error beta equal to probability fail to reject H 0 given H 0 is false. Now this is very interesting one and very important one also. Because if you accept H 0 when it is false, and when then you take many decisions which ultimately leads to very erroneous decisions which will be very costly also. So, another concept we will use here is power. What is power? Power is the 1 minus beta. This is this is nothing but probability of rejecting H 0, when H 0 is false.

So, we will now show you the sample size calculation using this hypothesis testing framework. So, primarily we will be using either beta or 1 minus beta. And we will see that beta or 1 minus beta is a function of sample size n, and then for a particular beta or 1 minus beta. We what we want to achieve, then what will be the sample size n. So, that that particular beta or 1 minus beta can be achieved.

(Refer Slide Time: 15:18)



So, see the situation here. Situation is if H 0 is true, this is the left-hand side normal distribution, we are assuming z distribution here. So, if H 0 is false, that actually mu is not equal to mu 0. That mean mu is mu 0 plus delta; then this under this condition H 1 is true. So, then the underlying distribution is the right-hand side one.

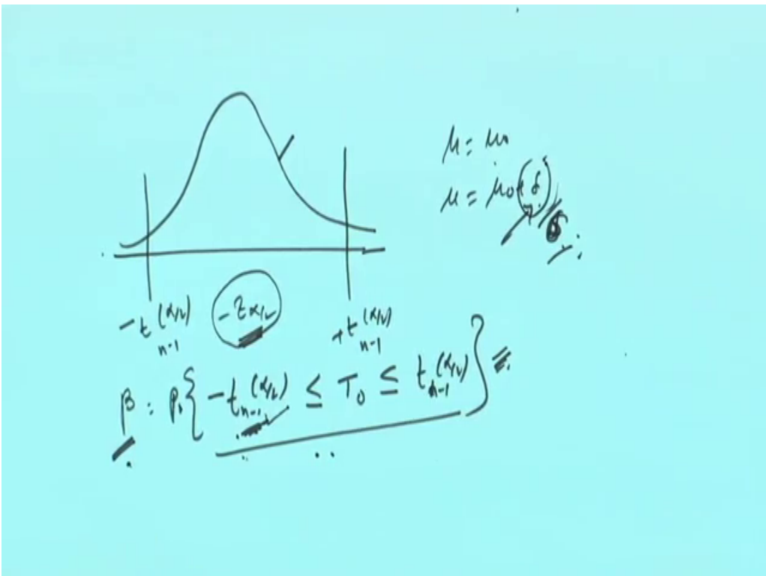
So, what how do we test? We test calculate that z 0 which is x x bar or y bar here we are reading x bar minus mu 0 by sigma by root n we are considering x is the x is the variables of importance. Or we can write z 0 equal to all through we have written this way mu 0 by sigma by root n. So, if mu equal to mu 0, then what will be the expected value of z 0, under null hypothesis. It will be expected value of either x bar minus mu 0 by sigma by root n. That is what you want to calculate. So, this is nothing but basically expected value of your; if I just do little bit of manipulation, we will get what we are getting mu equal to mu 0 plus delta.

So, mu 0 equal to mu minus delta, then then this will ultimately leads to expected value of x bar minus mu plus delta by sigma by root n. And then x bar minus mu by sigma by root n, this this will be 0, that expected value of this will be 0, but expected value of this will be there. That means expected value of delta by sigma by root n which is nothing but delta root n by sigma. So, if H 1 is true, then instead of this quantity x bar minus mu 0 by sigma by root n. Or expected value of y bar minus mu 0 by sigma by root n which one you take you take y is the random variable or x is random variable depending on this will not become 0, these become delta by sigma root n. That is what is given here. So, that mean the second distribution is normal distribution with mean sigma s delta root n by sigma and variance is 1.

Then what is the beta? Beta is your; you go back what is beta? Fail to reject H 0 H 0 is false. So, that mean what we are writing this is the distribution under H 0. So, but h, but H 0 is not true H 0 it is H 1 is true. So, any value falling here what will happen it will be although H 0 is not true, but we will accept it. Then what is the probability that H 0 is not true, but it is accepted is the overlapping portion between these and this; because minus z alpha minus z alpha by 2 plus z alpha that is the reason which is the acceptable reason.

So, the overlapping portion under H 1 true is this sided portion. This you will find out this probability, and then you find out the probability will be like this. This is the equation probability. Now in case that is not your sigma is not known, you will a sample standard deviation, then this quantity x bar minus mu 0 by sigma by root n this quantity follows; t distribution and your beta will be calculated using this equation. That beta will be probability minus t alpha by 2 n minus 1 to t alpha by 2 n minus 1, that mean what happened in this case?

(Refer Slide Time: 20:14)

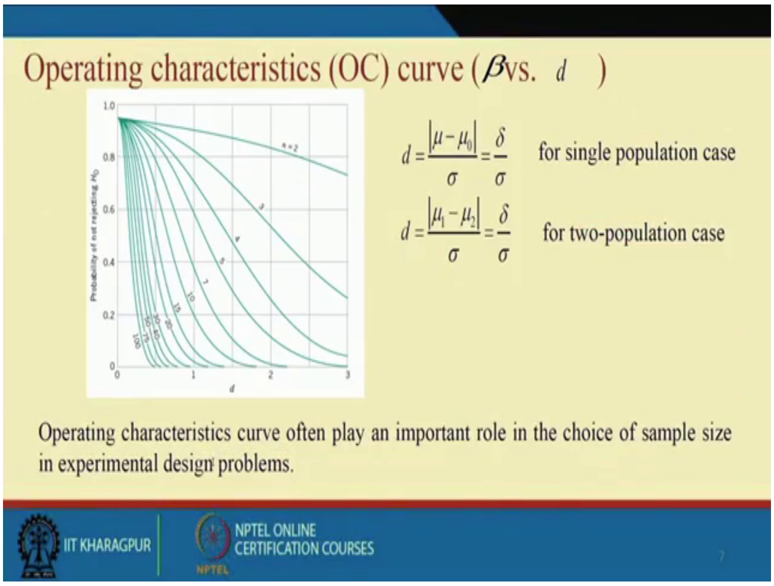


In this case these done several time. So, this will be minus t alpha by 2 n minus 1 and this will be plus t alpha by 2 n minus 1.

So, beta will be your probability that minus t n minus 1 alpha by 2 less than equal to t 0, less than equal to t alpha n minus 1 alpha by 2 this; however, this cannot be because she here what happened n is there, but if when z is there if it z distributed, then will write z alpha by 2 minus z alpha by 2 which is independent of which is divide of n. So, you we require a curve, like earlier I shown you that t distribution time you use a curve; so this kind of curve. So, beta a is dependent on n; so now, beta versus n. So, as such what happened, we find out that from these equations that delta by sigma and beta they are dependents. So, if you plot beta versus delta by sigma for a particular sample size, then this curve is known as operating characteristic curve or OC curve.

And or you can produce B beta versus n also that is also OC curve. So, we will see that 2 curves, one is beta versus delta by sigma. Another one is beta versus n, or we will see 1 minus beta versus n, all are basically operating characteristic curve. The beauty of this curve is that given value of beta, and given a value of for a particular delta, you will be in a position to calculate what should be the sample size. And other way around given delta and sample size what will be the beta? So, those alternatively you can calculate.

(Refer Slide Time: 22:33)

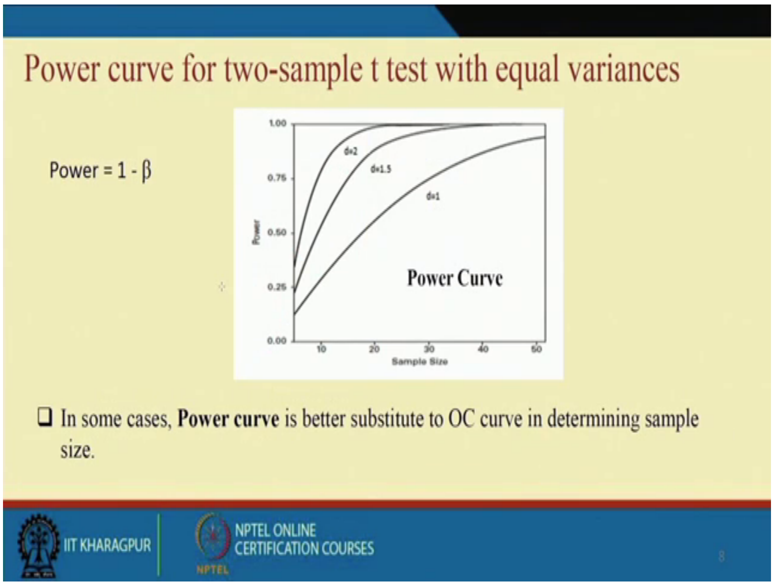


So now let us see that on operating characteristic curve where beta versus d. D is nothing but delta by sigma, which is a basically sigma waited delta. Delta is what? That mu h mu h not m 0 mu is mu 0 plus delta that is the mean shift. What is the shift? In mean that is delta.

So, these when it is divided by it is standard deviation by standard deviation sigma. Then we are creating a quantity called d, now probability of rej not rejecting H 0. When H 0 is false, that is what is your beta versus d you will be getting for different sample size it is like this, the curve is like this. So, you see if your sample size increases, then for a when a small d, you are you are able to reject it. So, ifs because a operating characteristics curve will give you the behavior with reference to beta and d, and also for different kinds of n. So, probability of not rejecting H 0 beta is very high suppose n equal 2, even though d is very large, but then also what happened? The probability of rejecting H 0 probability of not rejecting H 0 is very high beta average very high.

So, if you increase these you find out that the probability of not rejecting H 0 is reducing.

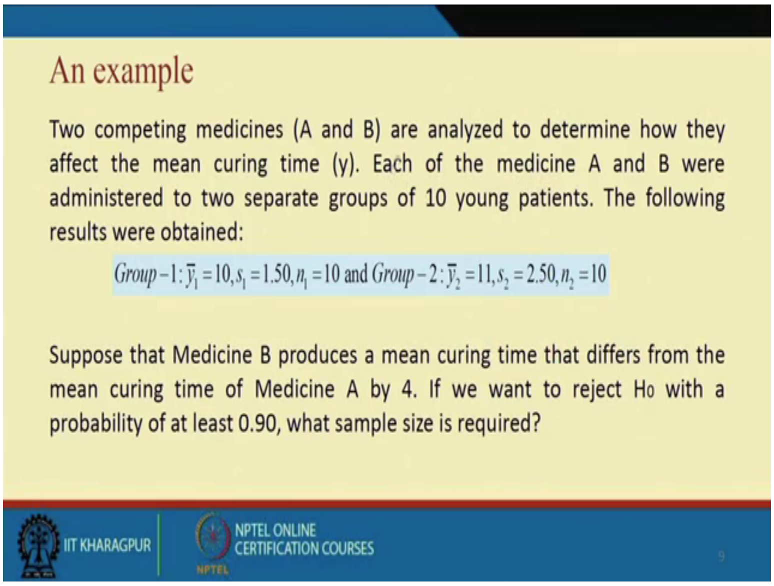
(Refer Slide Time: 24:25)



So, only use this power curve, what is the power curve? Power curve here instead of beta we are using 1 minus beta versus sample size for different value of d different value of d. So, if you know d and you basically fix a beta, then from this curve you know the sample size. And here, suppose if you know beta, and you know d and fix a value of beta, then you or fix a you want to see that the for a particular sample size what will be the beta that is also very impossible suppose I know d equal to 1, my sample size is suppose let it be 7. So, one and sample size 7, what will be the probability of not rejecting H 0, this you project to left hand side it will be like this. This value may be 0.38.

In power curve what happened? We here what we are saying that power is the is 1 minus beta, that mean probability of rejecting H 0 when H 0 is false. Suppose our d value is d value is 1. And then if our we want it to be that 80 percent power should be there or 0.1 minus, the 0.8, then you may find out that it t require around 50 number of observe a n observation to be collected that is what is your sample size. Now what I will do? I will give you one example.

(Refer Slide Time: 26:03)

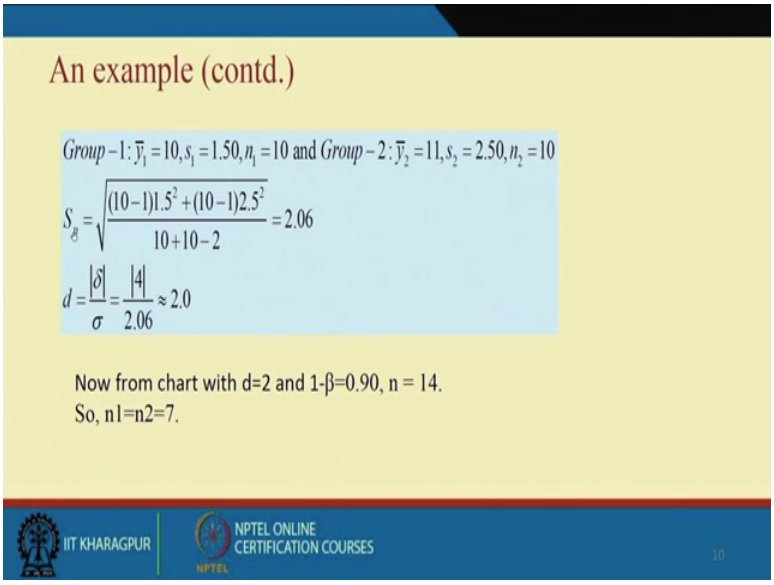


Two competing medicine A and B are analyzed to determine how they affect the mean curing time. Each of the medicine A and B were administered 2 separate group of 10 young patient the following results were obtained.

So, we are we are considering here that the 10 young patient, these 2 groups are similar in all other counts except, they are administered 2 different competing medicines. Suppose, the medicine B produces a mean curing time that differs from the mean curing time of medicine a by 4 unites. And if we want to reject H 0 with a probability of at least 0.90 what is the sample size required? What are the things given? Y 1 bar s 1 y 2 bar and n 1 and y 2 bar s s 2 and n 2 is given. In addition, what is given that the difference each 4 not 0. And in addition, what is given that we want a probability, we want to reject H 0 with a probability of at least 0.90.

So, what is the sample size that is required? We want to know the n whether the n 1 10 and n 2 10 is sufficient or not.

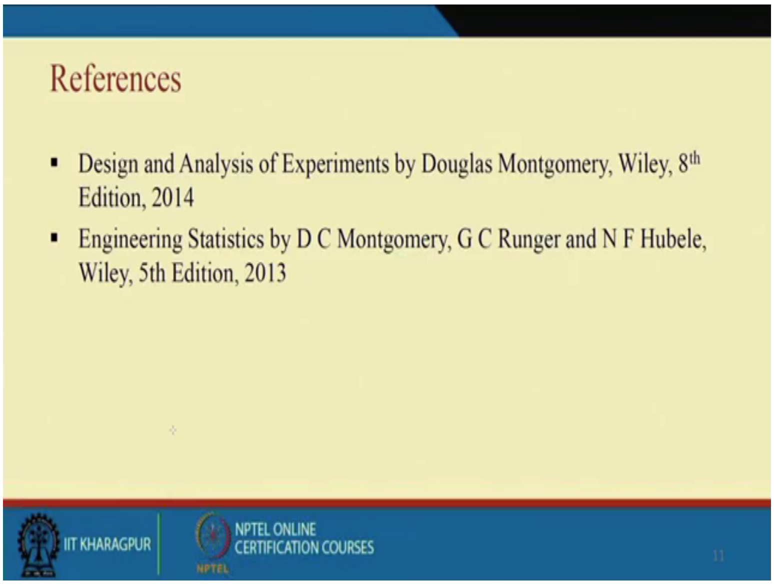
(Refer Slide Time: 27:19)



So, you first you calculate s p which is 2.06, then you find out delta. Delta is here almost 2. Now from the chart as beta is 0.90 beta 1 minus beta is 0.90, because quality of reject H 0 and H 0 is false. That is 1 minus beta 0.90 or d equal to 2, what value here getting from this? So, d 2 and 0.90 is 1 minus power is 0.90. So, if you just what you do if you take this is 0.90. Take a 5 under line here, it meets here, and then draw a particular line it will be somewhere here whose value is around 14.

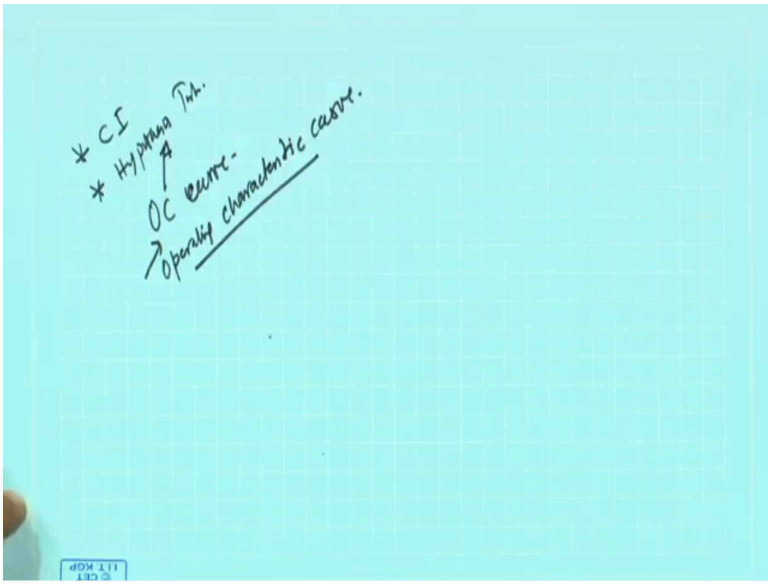
So, that mean n should be 14. So now, we have 2 different two populations. So, it will be for each population is each sample it will be n by 2; so n 1 equal to n 2 equal to 7. Roughly n by 2 we will give you the accurate results.

(Refer Slide Time: 28:28)



So, this is what the sample size determination. So, you can use the confidence interval approach or you can use hypothesis testing approach hypothesis testing approach.

(Refer Slide Time: 28:36)



And a in hypothesis testing approach the important concept what you run this OC curve; that is, operating characteristic curve. An OC curve is develop for different combination like it will be beta versus the shift mean shift. Or for given n sample size beta versus that n for if a given shift power versus that mean shift given an power versus your n given mean shift very, very popular concept. I hope you will be when you conduct experiment you will be in a position to know what should be the appropriate sample size, when you are conducting experiment involving one population or two population cases.

Thank you very much.